Deep Learning based approach to quantification of PET tracer uptake in small tumors*

Laura Dal Toso[1], Elisabeth Pfaehler[2], Ronald Boellaard[2,3], Julia A. Schnabel[1], and Paul K. Marsden[1]

[1] School of Biomedical Engineering & Imaging Sciences, King's College London, UK
[2] Department of Nuclear Medicine and Molecular Imaging, University of Groningen University Medical Center Groningen, Groningen, The Netherlands
[3] Department of Radiology & Nuclear Medicine, Amsterdam University Medical Centers, Amsterdam, The Netherlands
laura.dal_toso@kcl.ac.uk

Abstract. In Positron Emission Tomography (PET), quantification of tumor radiotracer uptake is mainly performed using standardised uptake value and related methods. However, the accuracy of these metrics is limited by the poor spatial resolution and noise properties of PET images. Therefore, there is a great need for new methods that allow for accurate and reproducible quantification of tumor radiotracer uptake, particularly for small regions. In this work, we propose a deep learning approach to improve quantification of PET tracer uptake in small tumors using a 3D convolutional neural network. The network was trained on simulated images that present 3D shapes with typical tumor tracer uptake distributions ('ground truth distributions'), and the corresponding set of simulated PET images. The network was tested on unseen simulated PET images and was shown to robustly estimate the original radiotracer uptake, yielding improved images both in terms of shape and activity distribution. The same network was successful when applied to 3D tumors acquired from physical phantom PET scans.

Keywords: Convolutional neural network · PET · Quantification · Reconstruction

1 Introduction

1.1 Positron Emission Tomography

Positron Emission Tomography (PET) is widely used in clinical oncology for the evaluation of lesion malignancy, staging and for monitoring the tumor response to treatment [11]. In the clinical routine, images are often interpreted by visual inspection, together with semi-quantitative measurements of tumor radiotracer

* This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 764458.

uptake such as standardized uptake value (SUV) and related metrics. The SUV is defined as follows:

$$SUV = \frac{activity\ concentration\ in\ ROI}{average\ activity\ concentration\ in\ whole\ body} \quad (1)$$

Two common ways of reporting SUV are SUV_{max} and SUV_{mean}. SUV_{max} represents the highest voxel value in a region of interest (ROI). This measurement is insensitive to the tumor boundary definition but it is very susceptible to noise. SUV_{mean} is an average SUV calculated over voxels in a boundary ROI. As a result, it is less sensitive to noise but it is dependent on the ROI definition and it typically has a lower value than SUV_{max}. An alternative metric is SUV_{peak} which is an average SUV calculated inside a small ROI, usually a 1 ml spherical volume, containing the pixel with maximum intensity [13]. SUV_{peak} is less affected by noise but depends on the ROI's shape, size and location. It is complicated to accurately quantify tumor uptake in PET images due mainly to poor spatial resolution, typically 5 mm FWHM, and noise [1,12]. Important advances have recently been made in the development of techniques such as tumor segmentation and image reconstruction, but there is a great need for new accurate and reproducible quantification methods and that can be easily integrated in clinical and clinical research settings [2].

1.2 Deep Learning in PET imaging

In recent years deep learning techniques have been massively applied to medical imaging. These approaches have been extremely successful in performing tasks such as segmentation, classification, automatic detection and, to a lesser extent so far, image reconstruction [7,8]. Convolutional neural networks (CNNs) have been successfully applied to PET images to perform denoising [5], lesion detection and lesion segmentation [3], as well as image reconstruction [4]. Even though a lot of progress has been made in these areas, there are only few deep learning applications explicitly aimed at improving quantification in PET imaging. In this work we present a deep learning approach with the aim of more accurately quantifying tumor radiotracer uptake in PET studies.

2 Materials and Method

One of the main obstacles that hamper the application of deep learning to PET imaging is the lack of large labelled image datasets, that are needed to train the networks. Because of the difficulty in obtaining ground truth radiotracer distribution data, in this work a simulation algorithm was developed to generate synthetic datasets that were used to train and test the network. A 3D CNN was trained on simulated PET images and on the corresponding ground truth radiotracer distributions. The 3D CNN learned the relationship between the two sets of images and, when presented with an unseen set of simulated PET images, it restored an estimate of the true radioactivity distribution. The proposed method, illustrated in Fig. 1, will be described in the following paragraphs.

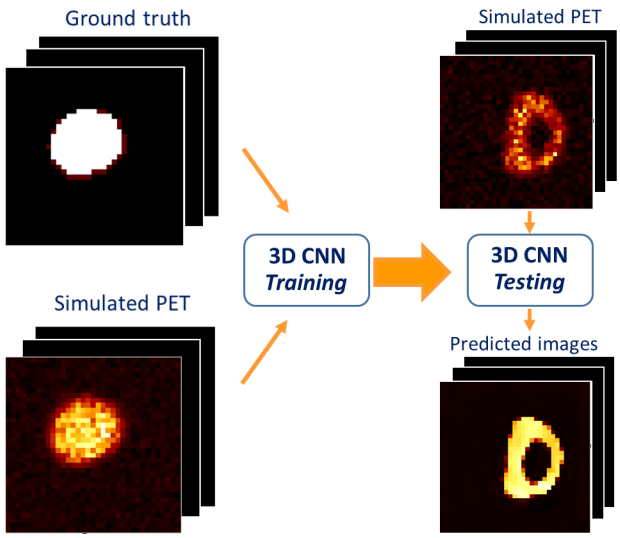


Fig. 1. A 3D CNN is trained on ground truth images and simulated PET images. The network is tested on an unseen set of simulated PET and it yields a prediction of the corresponding ground truth.

2.1 Generation of 3D shapes and radionuclide distribution

Two sets of data, henceforth called ground truth images and simulated PET images, were generated using a simulation algorithm. The set of ground truth images was composed of:

- subset 1: 1400 warped spheres filled with uniform activity
- subset 2: 1400 warped spheres divided in two halves. Different values of uniform activity were assigned to each volume, with a ratio of 3:1
- subset 3: 1400 hollow warped spheres. Different values of uniform activity were assigned to each volume, the activity in the inner part being one third of the activity in the outer part

The background of each image was set to $1/10^{th}$ of the maximum activity. These specific patterns of radiotracer distribution were chosen to simulate realistic heterogeneous tumor uptake distributions [10]. The simulated images were made of $35 \times 35 \times 40$ voxels and the voxel size was set to $3.18 \times 3.18 \times 2.00$ mm^3. In each subset the radius of the 3D shapes before the warping process spanned 2 to 12 voxels (6 to 36 mm) and the activity concentrations spanned 2 kBq/ml to 50 kBq/ml. After generating the ground truth images, the corresponding simulated PET images were produced by applying Gaussian convolution (FWHM varying between 4 and 6mm as described below), to simulate the effects of the point spread function (PSF) of the acquisition and reconstruction system and by adding noise. The noise values, assigned voxel by voxel, were drawn from a Gaussian distribution with $\sigma = k * \sqrt{N}$, where N is the number of counts in each voxel and k is a constant, set such that the noise level is equivalent to the noise observed in real PET data.

2.2 Network Architecture

The 3D network used in this work, was composed of five convolutional layers, each with 32 filters and $3 \times 3 \times 3$ filter size. Each convolutional layer was followed by a batch normalization layer to stabilize and accelerate the network training [6]. Finally, a fully connected layer with one hidden unit was used to obtain the output images. ReLu activation functions were assigned to the convolutional layers and a linear activation was used for the fully connected layer. The loss function was a mean squared error function, calculated between the predicted images and the ground truth images. The optimizer used to minimize the loss function during training was RMSprop. We used the Keras Framework with Tensorflow backend to implement the network, and training was performed on a NVIDIA Quadro M1200 GPU. Due to the limited available memory on the GPU, a batch size of 26 was used. The validation loss was monitored during the training process. The learning rate was set to the default value 0.001.

2.3 Testing the procedure

Images were visually inspected, using a software tool for multimodality medical image analysis (AMIDE) [9], as well as quantitatively assessed. When the ground truth was available, the predicted images were assessed by calculating the mean recovery coefficient RC_{mean}, defined as a ratio of the mean intensity value $\bar{I}$ of each prediction and its corresponding ground truth:

$$RC_{mean} = \frac{\bar{I}_{prediction}}{\bar{I}_{ground\ truth}} \tag{2}$$

The mean value for a given image was calculated over the voxels that exceeded a threshold of 50% of the maximum intensity. The mean structural similarity (MSSIM) [14] was also calculated to evaluate the similarity between the predicted images and the ground truth images. Three sets of experiments were performed to assess the performance of the procedure and the impact of different parameters on the CNN's predictions.

Normalization
When deep networks are used to perform tasks like segmentation or classification, it is common practice to normalize the input data to make the training faster. In our case the main goal is to improve quantification, so a study was performed to verify that absolute values are preserved during the normalization process. The dataset generated for this experiment, called original dataset, was made of 4200 images, divided into three subsets as described in section 2.1. The simulated PET images corresponding to the first, second and third subset were produced using a Gaussian function with FWHM = [4, 4, 4] mm, FWHM = [5, 5, 5] mm and FWHM = [6, 6, 6] mm respectively. To asses the impact of normalization, the original dataset was normalized by calculating the global maximum within the total set of simulated PET images and the total set of ground truth images,

and by dividing each voxel in all images by the global maximum. By using this method, a single scaling factor was used for all images so it was possible to easily rescale the predicted images to the original units. The CNN was separately trained on 3460 images belonging to the non-normalized dataset and on 3460 images taken from normalized dataset and the results were compared.

Different spatial resolutions

One of the aims of this work is to generate a model that can be easily applied in the clinical routine, so ideally it should not depend on the properties (in particular the PSF) of any specific scanner. At first, we trained and tested the CNN on a dataset including simulated PET images generated using the same PSF. Then, the same CNN was tested on simulated PET images produced using different PSFs. The training dataset generated for this experiment was made of 1120 images, that consisted of warped spheres filled with uniform activity distributions. The simulated PET images in the training set were created using a Gaussian function with FWHM = [4, 4, 4] mm. The first test set (TS1) was formed of 280 images generated in the same way. A second test set (TS2) was then created, using Gaussian functions with FWHM = [6, 6, 6] mm to generate the simulated PET images. In this experiment the images used for training and testing were not normalized.

Physical phantom PET scans

After training and testing the CNN on simulated images, we have also tested the network on a small set of phantom data acquired on a PET scanner. In this case, the network was trained on the same training dataset used to test the effects of normalization, where the simulated PET images belonging to the first, second and third subset were produced using a Gaussian function with FWHM = [4, 4, 4] mm, FWHM = [5, 5, 5] mm and FWHM = [6, 6, 6] mm respectively. The network was trained on 80% of the data belonging to the combined datasets made of three subsets, and tested on the remaining 20%. A fraction of 20% of the training set was used for validation. Then the network was tested on 3D patches extracted from real phantom PET scans. The physical phantom had the same size of a NEMA NU 2-2012 IQ phantom, that has a shape similar to a torso. Three 3D printed inserts simulating heterogeneous uptake distributions and realistic tumor shapes [10] were placed in the phantom at equal distances. The first tumor insert (T1) had a volume of 46.00 ml and was filled with an activity solution of 19.49 kBq/ml. The second tumor (T2) was divided into two parts: the upper part (10.75 ml) filled with an activity solution of 10.94 kBq/ml and the lower part (13.12 ml) filled with an activity solution of 19.49 kBq/ml. The third tumor (T3) was hollow, the outer part (65.35 ml) filled with an activity solution of 19.49 kBq/ml and the inner core (7.80 ml) filled with non-radioactive water. The background compartment of the NEMA IQ phantom was filled with an activity solution of 1.94 kBq/ml. The phantom was scanned on a PET/ CT system (Biograph mCT-40 PET/CT, Siemens, Knoxville, TN, USA) and the scans were acquired as list-mode data. The data were reconstructed to obtain

a frame of 300 s, that is comparable to the scan time used in the clinic for patients, using an iterative ordered subset expectation maximization (OSEM) algorithm (3 iterations, 24 subsets), time-of-flight (TOF) iterative reconstruction (3 iterations, 21 subsets) and point spread function (PSF) modeling. The size of the 3D patches used to test the 3D CNN was $35 \times 35 \times 40$ voxels and each patch contained the image of one tumor insert.

3 Results

3.1 Normalization

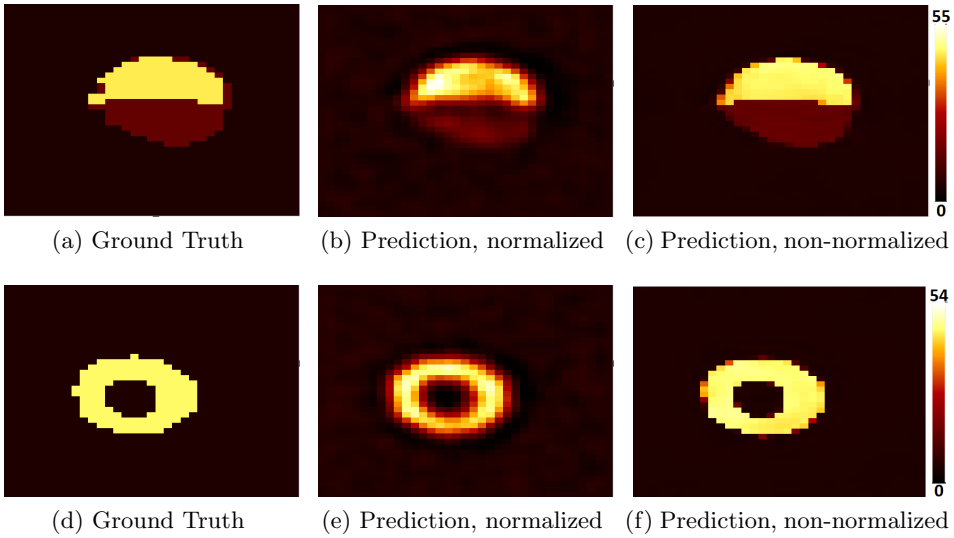
The predicted images obtained for two representative volumes are shown in Fig. 2. The first column shows a coronal section of the ground truth for each volume. The second and third column show the corresponding predictions obtained training and testing the CNN on the normalized dataset and on the non-normalized dataset respectively. The images in Fig. 2 (c) and Fig. 2 (f) are more similar to the ground truth: the edges of the active volumes are more clearly defined and the predicted intensities are closer to the ground truth activity distributions. This visual assessment is supported by the RC_{mean} values and by the mean structural similarity shown in Fig. 3. These graphs confirm that the images predicted by the CNN tested on the non-normalized dataset are overall characterized by higher MSSIM and RC_{mean} values, meaning that they are overall more similar to the ground truth. For this reason, non-normalized data have been used in the subsequent experiments.

3.2 Different spatial resolutions

In Fig. 4 three transverse views of two representative volumes belonging respectively to TS1 (top row) and TS2 (bottom row) are shown. The CNN yields better predictions when tested on TS1, in which the simulated PET images were produced using the same Gaussian function as in the training set. By visual comparison we can notice that the prediction in Fig. 4 (c) looks more similar to the ground truth in Fig. 4 (a) both in terms of shape and activity distribution. The prediction in Fig. 4 (f) has an overall lower activity and blurrier edges than the corresponding ground truth in Fig. 4 (d). The results obtained for 3D shapes smaller than 40 ml, presented in Fig. 5, show that the CNN better recovers the mean intensity in the warped spheres belonging to the first test set, which has the same PSF as the training data. This indicates the importance of matching the PSFs in the training set and in the testing set.

3.3 Physical phantom PET scans

A fraction of 20% of the simulated PET images belonging to the combined dataset (made of three subsets), that had not been used for training, was used at first to test the CNN. The results obtained in this case showed that the CNN

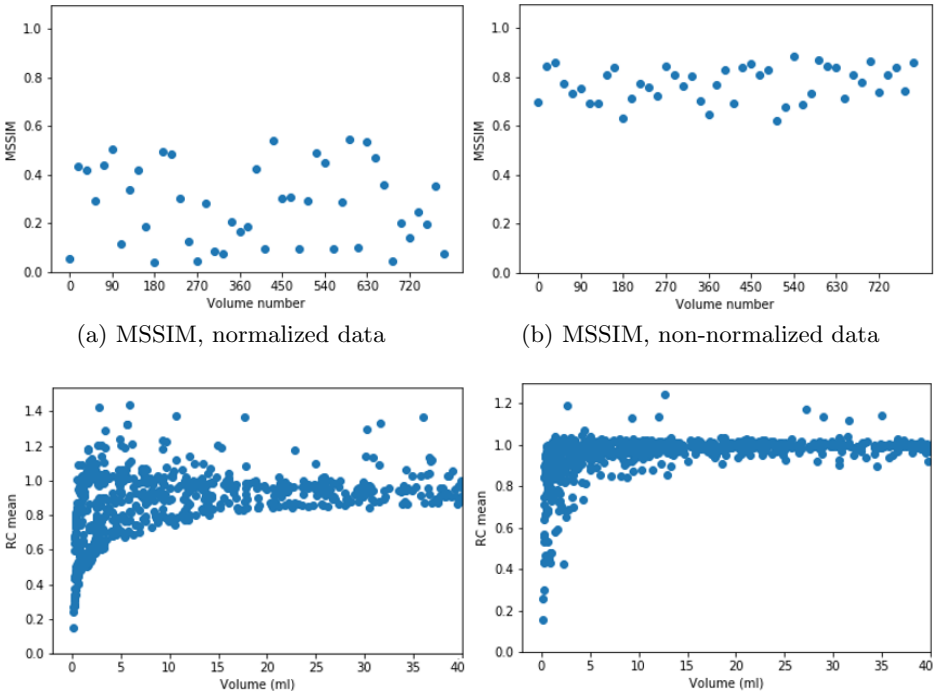


(a) Ground Truth (b) Prediction, normalized (c) Prediction, non-normalized

(d) Ground Truth (e) Prediction, normalized (f) Prediction, non-normalized

Fig. 2. Each row shows three coronal views, belonging to two representative volumes. The ground truth images are shown in the first column. The predicted images obtained testing the CNN on the normalized dataset, rescaled to the original units, are shown in the second column. The results obtained testing the CNN on the non-normalized dataset are shown in the third column. The network performs more effectively if the training data is not normalized. The intensities are expressed in kBq/ml.

could recover well the ground truth activity distributions and shapes. Then the same network was tested on phantom data, the results obtained testing the network on 3D images of the phantom inserts T1 and T2 are shown in Fig. 6. Three images, reconstructed using OSEM, PSF and PSF+TOF are shown for each tumor. Directly under each phantom image, the corresponding CNN's prediction is presented. The images yielded by the CNN are less noisy and the edges of the tumors are better defined. Due to the lack of ground truth images, in this case it was not possible to estimate the MSSIM and RC_{mean}. The quantification was performed estimating the maximum intensity voxel in each volume, a measurement that can be related to SUV_{max}. The maximum values extracted from the real phantom images and from the predicted images are presented in table 1. The ground truth maximum value is 19.49 kBq/ml for all tumor inserts. The predicted maximum values range from 18.18 kBq/ml to 23.98 kBq/ml, and are closer to the ground truth than the ones calculated for the real phantom scans. Although the CNN has a denoising effect, some noise is still present in the predicted images, which explains the variation observed in the predicted maximum values.



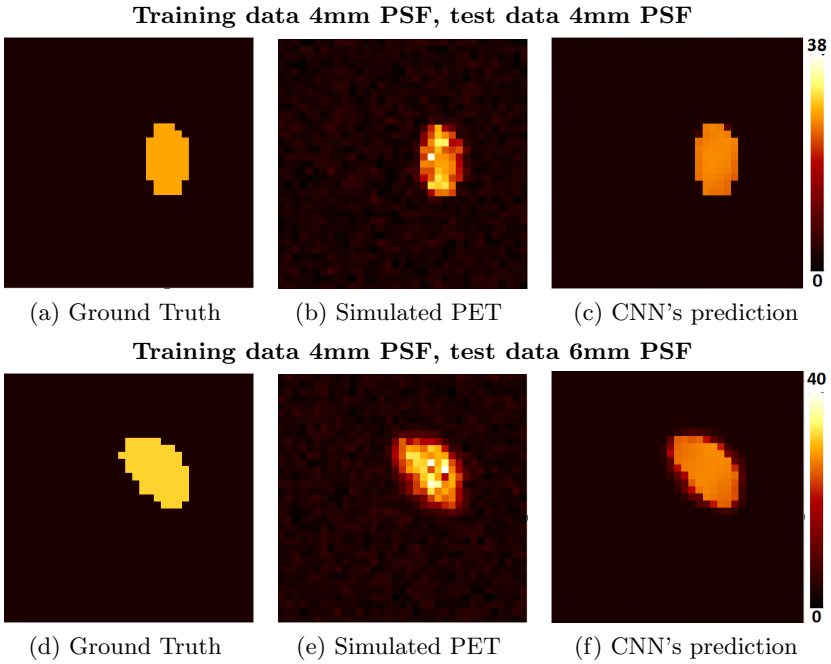
(a) MSSIM, normalized data

(b) MSSIM, non-normalized data

(c) RC_{mean}, normalized data

(d) RC_{mean}, non-normalized data

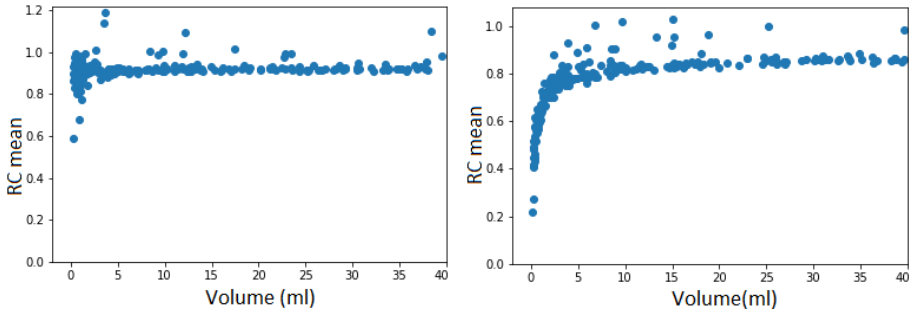
Fig. 3. Representation of MSSIM and RC_{mean} values, calculated for the predicted volumes belonging to the normalized dataset on the left and to the non-normalized dataset on the right. Only 45 representative MSSIM values are plotted to allow for a better visualization.



(a) Ground Truth (b) Simulated PET (c) CNN's prediction

(d) Ground Truth (e) Simulated PET (f) CNN's prediction

Fig. 4. Illustration of three transverse views of two representative volumes, belonging to TS1 (top row) and to TS2 (bottom row) respectively. The ground truth images are presented in the first column. The corresponding simulated PET images, generated using a PSF with FWHM = [4, 4, 4] mm and FWHM = [6, 6, 6] mm are shown in the second column, in Fig. (b) and (e) respectively. The predicted images yielded by the CNN are shown in the third column. The intensities are expressed in kBq/ml.



(a) Training data 4mm PSF; test data 4mm PSF

(b) Training data 4mm PSF; test data 6mm PSF

Fig. 5. Comparison between the mean recovery coefficients calculated for the predicted volumes obtained testing the CNN on TS1 (a), characterised by a PSF with FWHM = [4, 4, 4] mm, and on TS2 (b) characterised by a PSF with FWHM = [6, 6, 6] mm. A better recovery is obtained when PSFs are the same for training and test data.

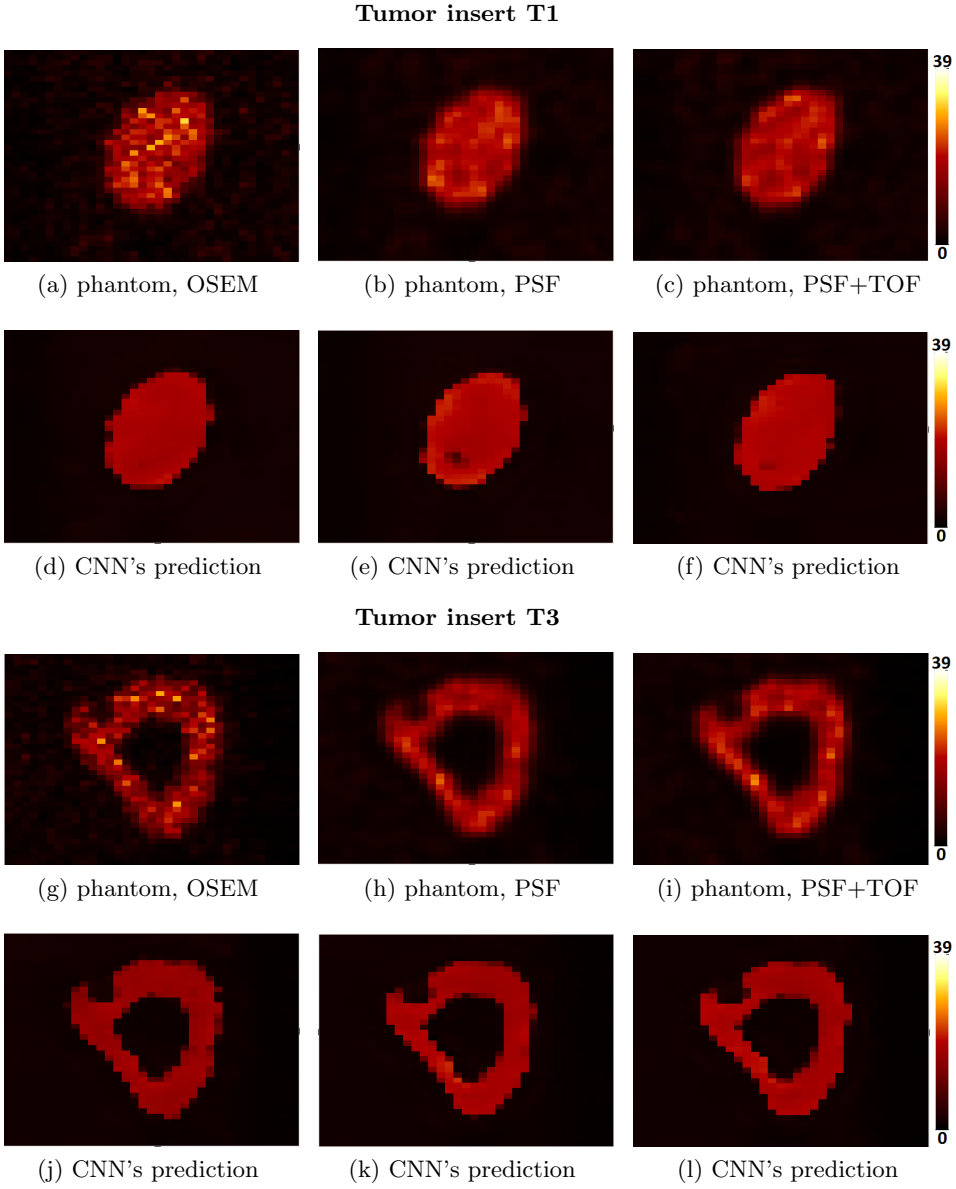


Fig. 6. The first and third row show three coronal views of the 3D patches extracted from phantom PET images, reconstructed using different algorithms (OSEM, PSF and PSF + TOF). Directly under each phantom image the corresponding CNN's prediction is presented. The intensities are expressed in kBq/ml.

	T1 phantom	T1 pred.	T2 phantom	T2 pred.	T3 phantom	T3 pred.
OSEM	34.77	19.19	31.79	18.18	38.78	21.72
PSF	26.65	20.90	24.15	23.98	27.19	22.30
PSF+TOF	25.34	21.76	25.35	20.89	27.56	21.08

Table 1. Maximum intensity values calculated for the real phantom scans and for the corresponding CNN's predicted images, expressed in kBq/ml.

4 Discussion and conclusion

In this paper, we have developed a deep learning approach to improve tumor radiotracer quantification in PET images. A simulation algorithm was implemented to generate the labelled datasets needed for training. The algorithm can very effectively recover a more accurate estimate of the original distribution from the simulated PET images. It has been demonstrated that, for our purpose, the network was performing better when trained on non-normalized data. The predictions obtained by training the network on non-normalised data had better defined edges around the active volumes and the predicted intensities were more similar to the ground truth activity distributions. Simulated PET images generated using three different Gaussian blurring functions were included in the training set and, when tested on a dataset including the same three Gaussian functions, the CNN was able to correctly recover the ground truth images. This suggests that the network's performance does not have a strong dependence on the scanner-specific PSF. Preliminary results from applying the methods, trained on simulated data, to real phantom PET data are very encouraging. The main limitations of this method are the very simplified simulation of PET-like images and the poor knowledge of actual tumor ground truth distributions. The results presented in this work show that the proposed method has the potential to improve quantification of tumor tracer uptake, overcoming the challenges due to the lack of large labelled image datasets. This new quantification method could also be used as a part of an end-to-end image reconstruction process. In future work, we plan to train the network on more realistic images, simulated using more sophisticated simulation methods like Monte Carlo and to include some real patient data to the study.

References

1. Bai, B., Bading, J., Conti, P.S.: Tumor quantification in clinical positron emission tomography. Theranostics (2013). https://doi.org/10.7150/thno.5629
2. Baizán, A.N., Puig, D.R., Segura, J.P.: Evolution of quantification methods in oncologic 18F-FDG PET studies. Revista Espanola de Medicina Nuclear e Imagen Molecular **37**(4), 203–204 (2018). https://doi.org/10.1016/j.remn.2018.06.001
3. Blanc-Durand, P., Gucht, A., Schaefer, N., Itti, E., O. Prior, J.: Automatic lesion detection and segmentation of 18f-fet pet in gliomas: A full 3d u-net convolutional neural network study. PLOS ONE **13**, e0195798 (04 2018). https://doi.org/10.1371/journal.pone.0195798

4. Gong, K., Guan, J., Kim, K., Zhang, X., Yang, J., Seo, Y., El Fakhri, G., Qi, J., Li, Q.: Iterative PET image reconstruction using convolutional neural network representation. IEEE Transactions on Medical Imaging **38**(3), 675–685 (mar 2019). https://doi.org/10.1109/TMI.2018.2869871
5. Gong, K., Guan, J., Liu, C.C., Qi, J.: PET Image Denoising Using a Deep Neural Network Through Fine Tuning. IEEE Transactions on Radiation and Plasma Medical Sciences **3**(2), 153–161 (oct 2019). https://doi.org/10.1109/trpms.2018.2877644
6. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37. pp. 448–456. ICML'15, JMLR.org (2015), http://dl.acm.org/citation.cfm?id=3045118.3045167
7. Kim, K., Wu, D., Gong, K., Dutta, J., Kim, J.H., Son, Y.D., Kim, H.K., El Fakhri, G., Li, Q.: Penalized pet reconstruction using deep learning prior and local linear fitting. IEEE Transactions on Medical Imaging (2018). https://doi.org/10.1109/TMI.2018.2832613
8. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis (2017). https://doi.org/10.1016/j.media.2017.07.005
9. Loening, A., Sam Gambhir, S.: Amide: a free software tool for multimodality medical image analysis. mol imaging 2: 131-137. Molecular imaging **2**, 131–7 (08 2003). https://doi.org/10.1162/153535003322556877
10. Pfaehler, E., Beukinga, R.J., de Jong, J.R., Slart, R.H., Slump, C.H., Dierckx, R.A., Boellaard, R.: Repeatability of 18f-fdg pet radiomic features: A phantom study to explore sensitivity to image reconstruction settings, noise, and delineation method. Medical Physics (2019). https://doi.org/10.1002/mp.13322
11. Strauss, L.G., Conti, P.S.: The Applications of PET in Clinical Oncology. Journal of Nuclear Medicine **32**, 623–648 (1991)
12. van der Vos, C.S., Koopman, D., Rijnsdorp, S., Arends, A.J., Boellaard, R., van Dalen, J.A., Lubberink, M., Willemsen, A.T., Visser, E.P.: Quantification, improvement, and harmonization of small lesion detection with state-of-the-art PET. European Journal of Nuclear Medicine and Molecular Imaging **44**, 4–16 (2017). https://doi.org/10.1007/s00259-017-3727-z
13. Wahl, R.L., Jacene, H., Kasamon, Y., Lodge, M.A.: From recist to percist: Evolving considerations for pet response criteria in solid tumors. Journal of Nuclear Medicine **50**(Suppl 1), 1–50 (2009). https://doi.org/10.2967/jnumed.108.057307.From
14. Wang, Z., Bovik, A.C., Sheikh, H.R., Member, S., Simoncelli, E.P., Member, S.: Image Quality Assessment : From Error Visibility to Structural Similarity. IEEE Transactions on Image Processing **13**(4), 1–14 (2004). https://doi.org/10.1109/TIP.2003.819861